Data storage: the simple solutions

Data storage principles

Hiba ALQASIR

2021-2022

télécom saint-étienne
école d'ingénieurs / nouvelles technologies

Course objectives

- Provide an overview of storage technologies.
- Learn about the different file formats used in the professional world for storing data.
- Know how to define a database.
- Understand the implications of storing data in a database.

Course organization

- CM: 3 × 3h
- TP: 3 × 3h
- Individual work: 15h
- Evaluation: Exam 50%, TP 50%.
- Material: all slides and TP subjects will be posted on Mootse `https://mootse.telecom-st-etienne.fr/course/view.php?id=1070`

Topics

1. What is data storage and what solution do we have?
2. File formats, CSV, JSON, XML ...
3. Introduction to database (more details in Relational Databases course).

Who am I

- Master degree in Machine Learning and Data Mining (MLDM) from University of Jean Monnet (2017).
- Ph.D. in computer science (artificial intelligence) from University of Lyon (2020).
- Member of the education staff in Télécom Saint-Etienne.
- Member of “Image analysis and understanding” team in Hubert Curien Laboratory.

Who are you?

- How much do you know about data storage?
- What do you expect from this course?

What is Data?

Data

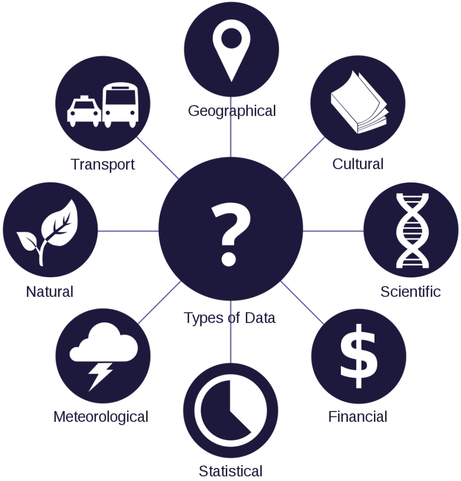


Image from fr.wikipedia.org

What is Storage?

Storage

Computer data storage

Technology for retaining data on a storage medium.

Image from ictexpertsluxembourg.lu

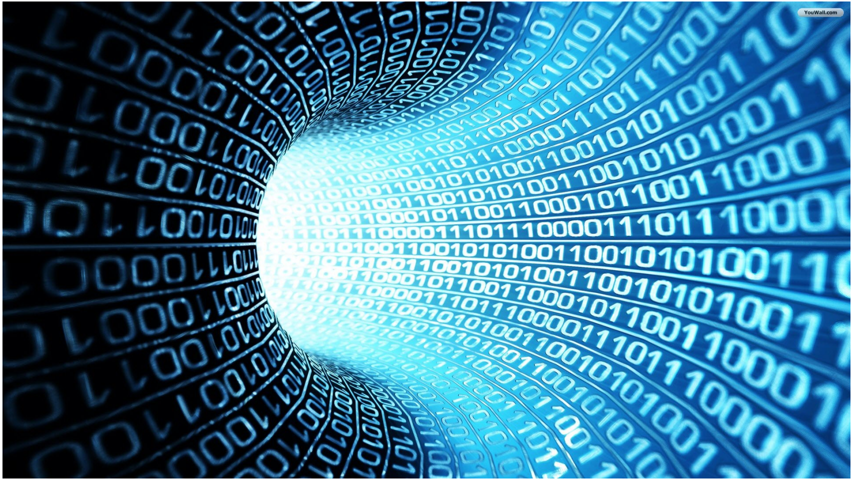
Storage medium

To place, keep and retrieve data.

- Technologies
- Devices
- Materials

- Price
- Speed
- Access mode
- Durability

How data is stored?



Bits

- **Binary digit**
- The **bit** is the smallest unit of storage
- Each **bit** is 1 or 0 (on or off)
- Anything with two separate states can store 1 bit
    - Electric charge = 0/1
    - Spots of North/South magnetism = 0/1
    - Flash of light = 0/1

Representing Data

- How many numbers can we represent with digits?

Representing Data

- How many numbers can we represent with digits?
    - 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Representing Data

- How many numbers can we represent with digits?
  - 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
- And after 9 ?

Representing Data

- How many numbers can we represent with digits?
  - 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
- And after 9 ?
  - We add new column and start again
  - 10, 11, 12 ... 99

Representing Data

- How many numbers can we represent with digits?
  - 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
- And after 9 ?
  - We add new column and start again
  - 10, 11, 12 ... 99
- And after 99 ?
  - We add another column!
  - 100, 101, 102 ... 999

Representing Data

- How many numbers can we represent with bits?

Representing Data

- How many numbers can we represent with bits?
  - 0, 1

Representing Data

- How many numbers can we represent with bits?
  - 0, 1
- And after 1 ?

Representing Data

- How many numbers can we represent with bits?
    - 0, 1
- And after 1 ?
    - We add new column and start again
    - 10, 11

Representing Data

- How many numbers can we represent with bits?
  - 0, 1
- And after 1 ?
  - We add new column and start again
  - 10, 11
- And after 11 ?
  - We add another column!
  - 100, 101, 110, 111

Representing Data

decimal	binary
0	0
1	1
2	10
3	11
4	100
5	101
6	110
7	111
8	1000

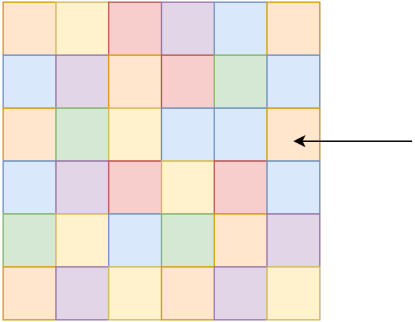
Everything is bits

BIG IDEA: Bits can represent anything!!

Numbers, characters, logical values, colors ...

Everything is bits

- American Standard Code for Information Interchange (ASCII)
  - ‘A’ has the ASCII code 65 (1000001 in binary).
  - ‘a’ has the ASCII code 97 (1100001 in binary).
- RGB images

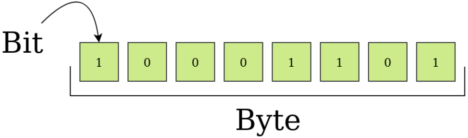


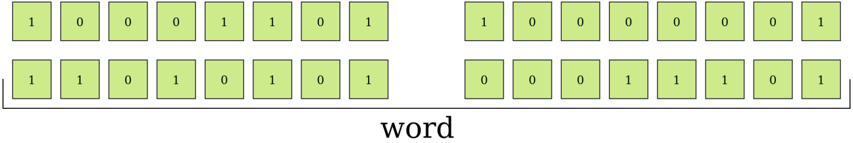
RGB (255, 230, 204)

R = 11111111
G = 11100110
B = 11001100

Bytes and Words

- Group 8 bits together into bytes
- Group 4 or 8 bytes together to make a word





MEMORIZE: N bits → 2^N combinations

Units of measurement for storage data

Decimal units

Name	Symbol	Value (base 10)	Value
byte	B	10^0	1 B
kilobyte	KB	10^3	1000 B
megabyte	MB	10^6	1000 KB
gigabyte	GB	10^9	1000 MB
terabyte	TB	10^{12}	1000 GB
petabyte	PB	10^{15}	1000 TB
exabyte	EB	10^{18}	1000 PB

Units of measurement for storage data

Binary units

Name	Symbol	Value (base 2)	Value
byte	B	2^0	1 B
kibibyte	KiB	2^{10}	1024 B
mebibyte	MiB	2^{20}	1024 KiB
gibibyte	GiB	2^{30}	1024 MiB
tebibyte	TiB	2^{40}	1024 GiB
pebibyte	PiB	2^{50}	1024 TiB
exbibyte	EiB	2^{60}	1024 PiB

Units of measurement for storage data

Binary units vs. decimal units

- The difference between 1 KB and 1 KiB is 2.4%.
  - $10^3 = 1000$
  - $2^{10} = 1024$
- The difference between 1 GB and 1 GiB is 7.4%.
  - $10^9 = 1000000000$
  - $2^{30} = 1073741824$

Exercise #1

You have 0.5 GB of images and 1635 KB of text files.
How much space do they take up overall in MB?

Exercise #2

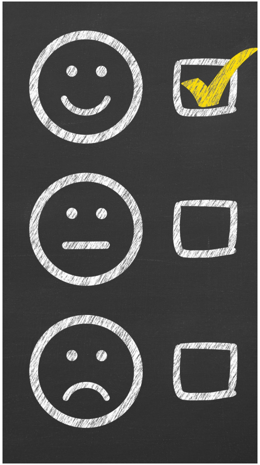
You have 450 images, each of them is 900 KB.
Will they all fit on your 2GB USB drive?

Exercise #3

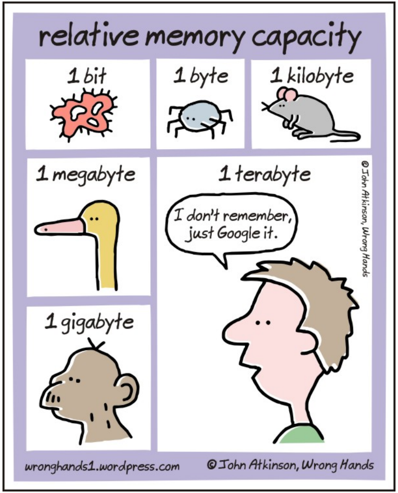
How many movies can be stored on a 1 TB drive?
If the size of each movie is 4 GB.

Characteristics of storage

- Capacity
- Performance
- Accessibility



Capacity



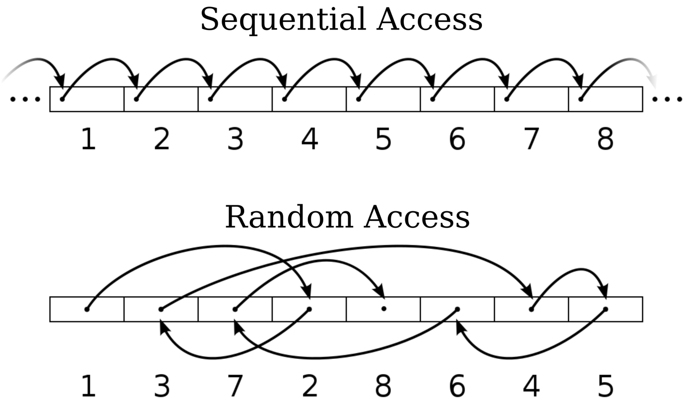
relative memory capacity
1 bit
1 byte
1 kilobyte
1 megabyte
1 terabyte
I don't remember, just Google it.
© John Atkinson, Wrong Hands
1 gigabyte
wronghands1.wordpress.com
© John Atkinson, Wrong Hands

Image from wronghands1.com

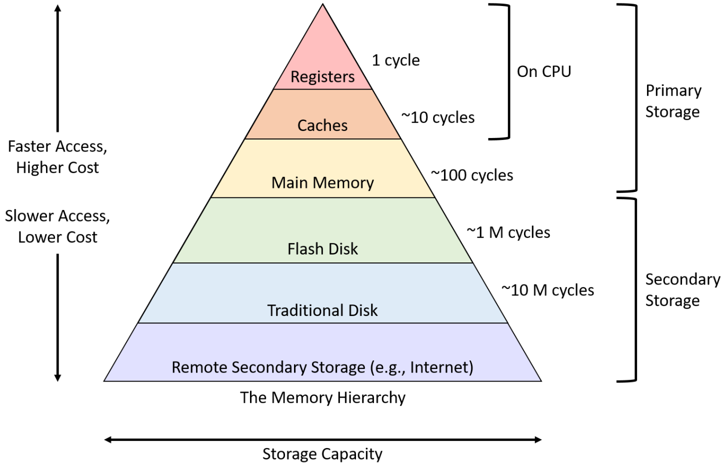
Performance

- Access time (Latency)
- Throughput
- Granularity
- Reliability

Accessibility



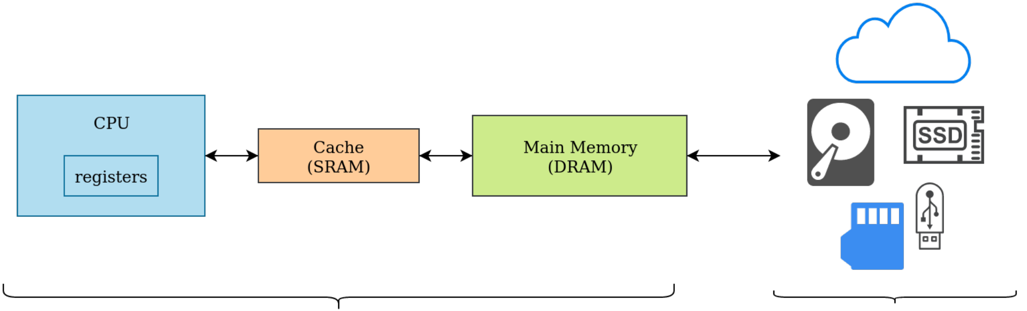
Hierarchy of storage



The Memory Hierarchy

Image from: diveintosystems.org/book/C11-MemHierarchy/mem_hierarchy.html

Hierarchy of storage



CPU

registers

Cache
(SRAM)

Main Memory
(DRAM)

Primary storage

(Volatile)

Secondary storage

(NonVolatile)

Ps: Nothing lasts forever.

Primary storage

- Cache memory (Static RAM)
- Main memory (Dynamic RAM)

Secondary storage

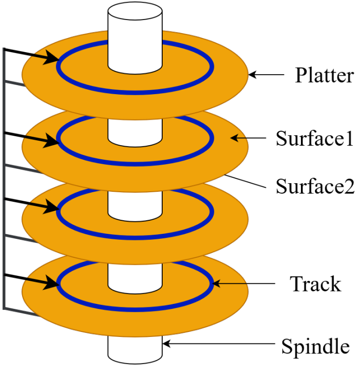
- Hard disk drive (HDD)
- Solid State Drive (SSD)
- Flash memory, CD, DVD, SD crads ...
- Cloud storage

Hard disk drive (HDD)

- Secondary storage
- Non-volatile storage
- Magnetic storage
- Highest-capacity in 2021 is 20 TB

Disk Geometry

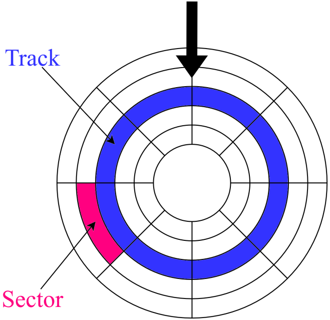
- The **disk** is made up of platters.
- Each **platter** has two surfaces.
- Each **surface** is made up of concentric rings called tracks.



Multi platter view

Disk Geometry

- Each **track** is made up of sectors.
- The size of a **sector** is usually 512 bytes.



Top view of single platter

Disk Capacity

Maximum number of bits that can be stored on the disk.

Capacity = #platters × #surface per platter × #tracks per surface × average #sectors per track × #bits per sector

Exercise #4

A hard disk with:

- 512 bytes per sector
- 1024 sector per track (on average)
- 2048 tracks per surface
- 1 surface per platter
- 5 platters

What is the total capacity of this disk?

Exercise #5

How many sectors per track (on average) are there in a disk has:

- 2048 tracks per surface
- 2 surfaces per platter
- 5 platters

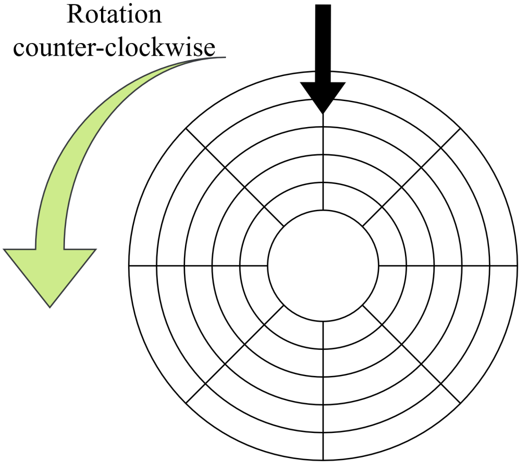
Knowing that its capacity = 0.5 TB.

The block

- The **block** is the input/output unit between secondary memory and primary memory.
- Block size is 512 bytes (1 sector) or 1024 bytes (2 sectors) ... or 8192 bytes (16 sectors).
- Any reading from or writing to disks is done in blocks.

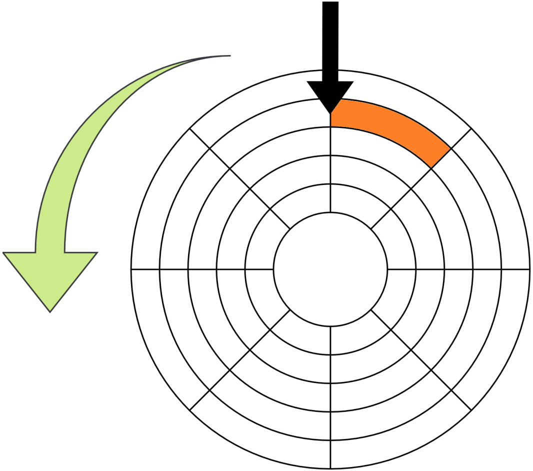
Disk Access

Block = 1 sector



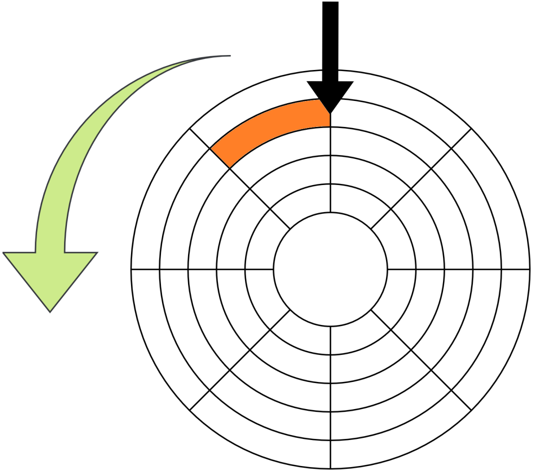
Disk Access

Before reading the orange block



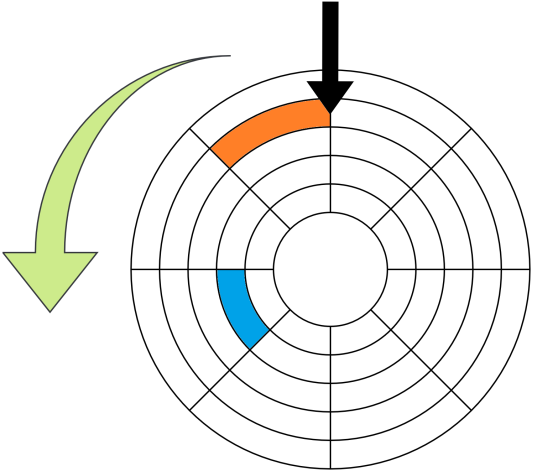
Disk Access

After reading the orange block



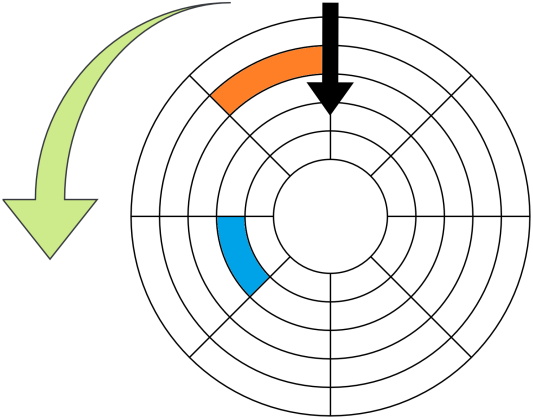
Disk Access

Request to read blue sector



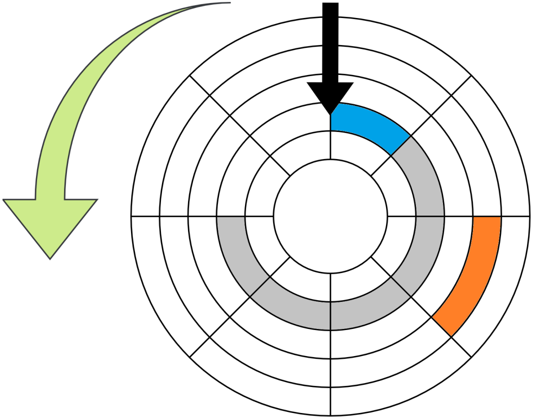
Disk Access

Seek to the track of the blue block



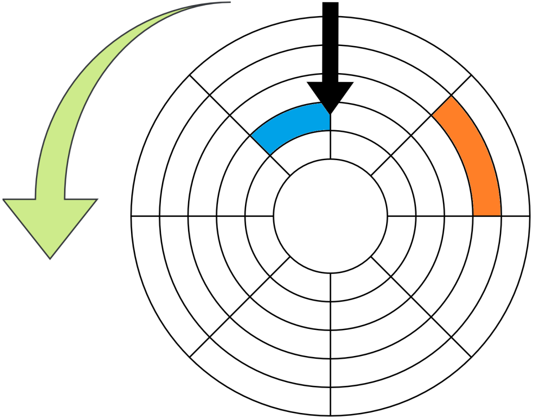
Disk Access

Rotation



Disk Access

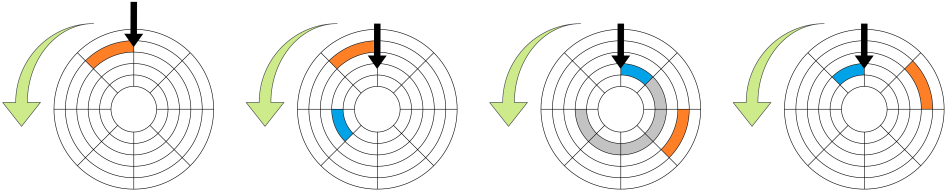
After reading the blue block



Disk Access Time

Average time to access a target block is estimated by:

- seek + rotation + transfer



data transfer seek rotation data transfer

Disk Access Time

- Seek time
  - Time to position the reading head on the track containing the block
  - 3-9 ms (defined by the manufacturer).
- Rotational latency
  - Time for the target block to arrive under the head
  - $1/2 \times 1/\text{rotation rate}$
- Transfer time
  - Time to read the target block
  - Depend on the transfer rate

Exercise #6

Given that:

- Rotation rate = 10000 RPM
- Transfer rate = 50 MB/sec.
- Average seek time = 5ms

How much is the access time to read 1024 bytes?

Solid-State Drive (SSD)

- Secondary storage
- Non-volatile storage
- Flash memory
- Highest-capacity in 2021 is 100 TB (40000$)

Cloud storage

Image from medium.com